JOURNAL OF ADVANCED ELECTRICAL AND ELECTRONIC ENGINEERING

Vol 2 Issue 1 (2026)

http://jaeee.org/

Synchronous Voice Action Dialogue Policies Coordinate Service Robot Turn Taking Control and Robot Motion Timing

Avery Bailey, Markus Neumann
Faculty of Technology, University of Oulu, Oulu, Finland

Abstract
The integration of physical actions with verbal communication remains a critical challenge in the development of socially competent service robots. Human conversation is inherently multimodal, relying on a complex synchronization of speech, gaze, gestures, and utilitarian actions to regulate turn-taking smoothly. However, traditional dialogue systems typically process speech and robotic actions sequentially or asynchronously, leading to unnatural delays, interruptions, and a degraded user experience. This paper proposes a comprehensive framework for synchronous voice-action dialogue policies aimed at optimizing service robot turn-taking control. By treating physical actions and verbal utterances as a unified output space, the proposed model enables robots to anticipate turn transitions and execute synchronized multimodal responses. The research introduces a continuous state-space representation that captures both linguistic context and spatial-temporal features of the interaction environment. Through extensive simulated and real-world human-robot interaction trials, this study evaluates the efficacy of synchronous policies against conventional modular approaches. The findings demonstrate significant improvements in interaction fluency, characterized by a reduction in overlapping speech, lower response latencies, and higher subjective ratings of robot sociability. This research provides a foundational architecture for next-generation service robots, emphasizing the necessity of embodied cognition in spoken dialogue systems and offering new perspectives on multimodal reinforcement learning in domestic and public service environments.

Keywords

Service Robots, Turn-Taking Control, Dialogue Policy, Multimodal Interaction.

INTRODUCTION

The deployment of service robots in dynamic human environments has accelerated significantly over the past decade. These autonomous entities are no longer confined to structured industrial settings but are increasingly tasked with navigating the complexities of domestic, healthcare, and hospitality domains. In such environments, effective communication is paramount. Traditional approaches to human-robot interaction have predominantly focused on spoken dialogue systems inherited from virtual assistants. These systems generally operate on a rigid, turn-based paradigm where the human speaks, the system processes the audio, and the system responds with synthesized speech. While adequate for disembodied agents, this asynchronous, voice-centric model is fundamentally insufficient for embodied service robots [1]. Human communication is intrinsically multimodal. It involves a continuous, synchronized exchange of verbal and non-verbal cues, including gaze direction, facial expressions, hand gestures, and body posture, all of which dynamically regulate the flow of conversation. The inability of current service robots to synchronize their physical actions with their verbal outputs during conversational turn-taking results in interactions that humans perceive as disjointed, hesitant, and socially awkward [2].

Turn-taking is the fundamental regulatory mechanism of human conversation, ensuring that information is exchanged efficiently with minimal overlap or disruptive gaps. In human-human interaction, turn transitions are negotiated seamlessly, often occurring within fractions of a second. This remarkable synchronization is achieved because participants do not wait for the semantic completion of a sentence to plan their response. Instead, they rely on predictive models driven by syntactic trajectories, prosodic intonation, and, crucially, physical action [3]. For example, a person might begin extending their hand to receive an object while simultaneously completing their verbal request. If a service robot waits until the human has entirely finished speaking before initiating its own voice response, and subsequently waits for the voice response to conclude before beginning its physical task execution, the interaction latency becomes highly noticeable and frustrating for the user [4]. The central premise of this research is that service robots must possess dialogue policies that generate voice and action synchronously, treating them as coupled dimensions of a single interactive turn. The development of synchronous voice-action dialogue policies introduces profound architectural and algorithmic challenges. Traditional dialogue managers rely on discrete state tracking, which maps user intents to specific verbal responses. Extending this mapping to include complex, continuous physical actions requires a fundamental reconceptualization of the dialogue state space [5]. The robot must maintain a real-time understanding of not only the linguistic context but also the spatial arrangement of the environment, the physical state of the user, and the kinematic constraints of its own body. Furthermore, the decision-making process must operate under stringent temporal constraints. Deciding when to take the floor, when to yield, when to backchannel, and what physical motions should accompany these conversational states requires a highly optimized policy. This paper comprehensively addresses the problem of synchronous turn-taking control by proposing a novel, unified reinforcement learning framework for service robots. The primary objective is to develop a dialogue policy that simultaneously plans verbal utterances and motor commands, thereby mimicking the overlapping and predictive nature of human conversational dynamics. To achieve this, we introduce an integrated state-representation architecture that fuses real-time natural language understanding with computer vision inputs, tracking user gaze and proxemics alongside dialogue history. This rich state representation feeds into a synchronous decision-making engine that outputs coordinated multi-channel responses. The contributions of this study are multifaceted. First, we provide an extensive theoretical analysis of the temporal misalignments in current human-robot interaction paradigms. Second, we design and implement a novel multimodal dialogue policy using advanced machine

learning techniques, specifically adapted for the temporal synchronization of speech synthesis and robotic actuation. Third, we conduct rigorous empirical evaluations involving human participants interacting with a physical service robot in structured collaborative tasks. The results are analyzed using a robust set of objective metrics, such as gap duration and overlapping speech ratios, alongside subjective measures of perceived naturalness and trust. Through this comprehensive investigation, we aim to bridge the persistent gap between conversational artificial intelligence and robotic control systems, paving the way for service robots that interact with human-like fluency and social intelligence.

2. Literature Review and Background

2.1 Multimodal Human-Robot Interaction

The field of human-robot interaction has long recognized the importance of multimodality, drawing heavily from psycholinguistics and communication theory. Early research established that non-verbal cues carry a substantial portion of the communicative payload in face-to-face interactions. In the context of robotics, studies have demonstrated that users respond more positively to robots that exhibit appropriate gaze behaviors and conversational gestures [6]. However, the initial integration of these modalities was highly scripted. Early socially assistive robots utilized finite state machines to trigger specific animations in conjunction with pre-recorded speech. While these systems demonstrated the value of embodied communication, they lacked the flexibility required for unconstrained interactions. As the field progressed, researchers began exploring more dynamic integration strategies. The concept of behavior realization engines emerged, aiming to automatically generate appropriate gestures based on the semantic content of the planned utterance. For instance, systems were developed to emphasize specific words through beat gestures or to indicate spatial relationships using deictic gestures [7]. Despite these advancements, the generation process remained fundamentally sequential. The dialogue manager would determine the verbal response, pass the text to a natural language generation module, and subsequently send the output to a behavior realization engine to attach synchronized animations. This pipeline approach inherently precludes the possibility of action driving the dialogue or of actions and speech being co-planned to optimize turn-taking transitions. The physical action is treated as an accessory to the speech rather than an equal component of the conversational turn [8]. Recent paradigms have attempted to leverage deep learning to bridge this gap, utilizing sequence-to-sequence models to map user inputs directly to multimodal outputs. While these end-to-end approaches show promise in virtual avatars, their application to physical robots is complicated by the realities of hardware execution [9]. A physical robot cannot instantaneously assume a pose; it must calculate inverse kinematics and execute trajectories over time. If a dialogue policy does not account for these execution latencies, the resulting interaction will suffer from severe synchronization failures, where the robot speaks about an action it has not yet completed or completes an action long after the relevant speech has concluded.

2.2 Turn-Taking Dynamics in Spoken Dialogue Systems

Turn-taking is a highly orchestrated process governed by implicit social rules. Sacks, Schegloff, and Jefferson provided the foundational framework for analyzing conversation, highlighting the minimal gaps and minimal overlaps that characterize natural transitions [10]. In spoken dialogue systems, replicating this efficiency has been a persistent challenge. Early voice assistants relied on simple silence thresholds to detect the end of a user's turn. If the user paused for more than a predetermined duration, the system assumed the turn was complete and initiated its response. This approach is highly error-prone, as humans frequently use intra-turn pauses for cognitive processing or breathing. Relying solely on silence leads to frequent interruptions by the system, severely degrading the user experience [11].

To overcome the limitations of silence-based endpointing, researchers introduced continuous turn-taking models. These models utilize acoustic and prosodic features, such as pitch contours and energy levels, combined with syntactic completion probabilities, to predict the end of a turn before it occurs [12]. By anticipating the transition, the system can begin processing its response earlier, thereby reducing the gap duration. Some advanced systems also incorporate backchanneling capabilities, allowing the agent to produce short conversational acknowledgments without taking the floor. Despite these improvements in verbal turn-taking, the translation of these models to embodied agents reveals significant shortcomings. When a robot is engaged in a physical task, such as handing an object to a user or navigating through a space, its physical state becomes a crucial determinant of turn-taking behavior. If a robot is currently out of reach, a user may naturally delay their verbal request or extend their turn until the robot is proximate. Conversely, a robot might use a physical action, such as establishing direct eye contact or shifting its posture, to signal its intention to take the floor [13]. Current dialogue policies largely ignore these physical turn-yielding and turn-holding cues, leading to a disconnect between the robot's conversational state and its physical embodiment.

2.3 Limitations of Current Asynchronous Models

The core limitation of contemporary service robot architectures is the asynchronous processing of voice and action. In a typical modular architecture, the dialogue management system and the robotic control system operate on separate asynchronous threads, communicating via high-level message passing. When a user issues a command that requires both a verbal response and a physical action, the dialogue manager typically issues the speech command immediately while simultaneously dispatching an action goal to the navigation or manipulation stack [14]. Because the physical action takes considerably longer to plan and execute than the speech synthesis, the robot often finishes speaking before it has even begun to move. Consider a scenario where a user asks a service robot to fetch a specific item from a table. Under an asynchronous model, the robot might immediately say the item is being fetched while remaining stationary as it calculates collision-free trajectories. Alternatively, if the system is programmed to wait for motion completion before speaking, it might remain completely silent and unresponsive for several seconds while moving, leaving the user uncertain if the command was understood. Neither of these outcomes aligns with human expectations [15]. To address these limitations, researchers have explored intermediate solutions, such as implementing conversational fillers. If the physical action is delayed, the dialogue manager might insert phrases indicating cognitive or physical effort to bridge the silence. While helpful, fillers do not solve the underlying synchronization problem. They are a patch over the symptoms of an asynchronous architecture. True interactional fluency requires a synchronous policy where the initiation, execution, and completion of speech and physical movement are continuously coordinated and co-adapted in real-time based on the unfolding dynamics of the interaction [16]. This necessitates a departure from traditional state-action mappings and requires the development of unified state spaces that inherently link communicative intent with motor control.

3. Methodology and System Architecture

3.1 System Overview

To overcome the limitations of asynchronous dialogue management, we propose a Synchronous Voice-Action Dialogue Policy architecture. The fundamental principle of this architecture is the consolidation of conversational decision-making and high-level motor planning into a single, cohesive reinforcement learning environment. Instead of maintaining independent states for the dialogue manager and the robotic execution system, our framework constructs a unified multimodal state representation. This state continuously updates at a high frequency, capturing the micro-dynamics of the user's speech, the user's physical posture, the robot's current kinematic

configuration, and the progression of the robot's synthesized speech buffer. The architecture comprises three primary subsystems: the Multimodal Perception Module, the Synchronous Policy Engine, and the Coordinated Execution Controller. The Multimodal Perception Module is responsible for aggregating sensor data. It processes audio streams to extract acoustic features, performs real-time automatic speech recognition to derive semantic intent, and utilizes RGB-D camera feeds to track user skeletal structures, head pose, and gaze vectors. The Synchronous Policy Engine is the core decision-making component. Formulated as a Markov Decision Process, it evaluates the unified state and selects an action from a joint voice-action space at every time step. Finally, the Coordinated Execution Controller translates the high-level policy decisions into low-level joint trajectories and audio output streams, ensuring that the timing constraints dictated by the policy are strictly maintained during physical realization.

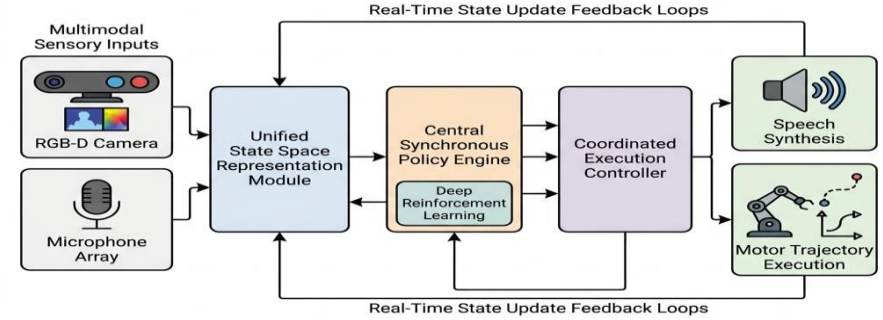


Figure 1: Comprehensive Schematic of Synchronous Voice

The design of the perception module specifically targets the latency issues inherent in sequential processing. Rather than waiting for a complete sentence to perform natural language understanding, the system utilizes incremental processing. Words are analyzed as they are recognized, allowing the dialogue state to reflect partial intents. Simultaneously, the visual processing pipeline maintains a continuous estimate of the user's focus of attention. If a user turns their head towards a specific object while beginning to speak, the state representation captures this spatial-temporal correlation immediately, enabling the policy engine to begin planning a physical orientation or manipulation action before the user finishes their utterance.

3.2 Synchronous Voice-Action State Space

The efficacy of a reinforcement learning policy is fundamentally dependent on the comprehensiveness and structure of its state space. In our framework, the state space must encapsulate all variables relevant to both conversational turn-taking and physical task execution. We define the multimodal state as a continuous vector composed of four distinct feature sets: linguistic context, acoustic-prosodic dynamics, visual-proxemic parameters, and internal robotic state. The linguistic context features encode the semantic meaning of the interaction up to the current time step. This includes the history of user intents, the entities mentioned, and the progression toward task completion. The acoustic-prosodic dynamics capture the non-lexical aspects of the user's speech, which are critical for turn-taking prediction. This subset includes real-time calculations of voice activity, pitch variations, and energy envelopes. A sudden drop in pitch and energy, for instance, serves as a strong indicator that the user is about to yield the turn.

Visual-proxemic parameters represent the physical relationship between the user and the robot, as well as the user's non-verbal communicative signals. This includes the relative distance and orientation between the two agents, the user's gaze target probability distribution, and the detection of specific manual gestures, such as pointing or hand-raising. These features allow the policy to understand when a user is physically engaging with the robot, even in the absence of speech.

Table 1: Description of Multimodal State Space Features

Feature Category	Specific Variables Tracked	Sampling Frequency	Representation Format
Linguistic Context	Incremental intent probability, Entity slots, Dialogue history vector	Text-driven	High-dimensional embedding
Acoustic Dynamics	Voice activity detection, Fundamental frequency, Signal energy	High frequency	Continuous scalar values
Visual-Proxemics	Interpersonal distance, User gaze vector, Gesture classification	Frame-rate	Normalized coordinates and probabilities
Internal State	End-effector position, Current speech buffer status, Action completion percentage	Real-time control loop	Kinematic joint angles and time-series flags

The internal robotic state is perhaps the most crucial addition for synchronous control. Traditional dialogue systems are largely unaware of their physical embodiment. Our state vector includes the precise coordinates of the robot's end-effectors, the velocity of its base, and the status of any ongoing physical actions. Furthermore, it tracks the state of the text-to-speech engine, noting how many syllables remain in the current utterance buffer. By including its own physical and verbal execution status in the state representation, the robot can make informed decisions about whether to interrupt its own movement, pause its speech to complete a complex physical manipulation, or smoothly overlap a new verbal response with the conclusion of an ongoing gesture.

3.3 Policy Learning Framework

The decision-making process is modeled as an infinite-horizon, discounted Markov Decision Process. The action space is inherently multidimensional, requiring the policy to select a combination of verbal and physical behaviors simultaneously. To manage the complexity of this joint space, we discretize the outputs into a set of high-level synchronous primitives. These primitives represent coordinated behaviors, such as initiating speech while maintaining gaze, uttering a filler word while beginning a reaching motion, or performing a deictic gesture synchronized with a specific semantic token. The core algorithm relies on a deep reinforcement learning architecture. The neural network takes the comprehensive state vector as input and processes it through a series of fully connected layers. To handle the temporal dependencies of conversation, we incorporate recurrent memory structures within the network. This allows the policy to retain information about previous turns and physical interactions over extended periods, which is vital for maintaining conversational coherence and avoiding repetitive behaviors. The reward function is the most critical component in shaping the desired synchronous turn-taking behavior. It is designed to penalize interactional friction while rewarding task efficiency and conversational fluency. The function comprises multiple weighted components. A negative reward is applied for excessive overlapping speech, penalizing the robot if it interrupts the user inappropriately. Similarly, a penalty is incurred for long conversational gaps, encouraging the robot to respond promptly. Crucially, the reward function includes a synchronization term that evaluates

the temporal alignment between the robot's physical actions and its speech. If the robot issues a verbal command related to an object but fails to orient toward or manipulate that object within a strict time window, a negative synchronization penalty is applied. Conversely, successful coordination, such as completing a hand-over action precisely as the accompanying spoken phrase ends, yields a substantial positive reward.

Code Listing 1: Synchronous Dialogue Policy Environment Initialization

```
class SynchronousPolicyEnv(gym.Env):
    def __init__(self, config_params):
        super(SynchronousPolicyEnv, self).__init__()
        self.linguistic_dim = config_params.get("linguistic_dim", 256)
        self.acoustic_dim = config_params.get("acoustic_dim", 12)
        self.visual_dim = config_params.get("visual_dim", 18)
        self.internal_dim = config_params.get("internal_dim", 24)
        total_state_dim = (self.linguistic_dim + self.acoustic_dim +
                           self.visual_dim + self.internal_dim)
        self.observation_space = spaces.Box(
            low=-1.0, high=1.0, shape=(total_state_dim,), dtype=np.float32
        )
            self.action_space = spaces.MultiDiscrete([
            config_params.get("num_verbal_actions", 50),
            config_params.get("num_physical_actions", 20),
            config_params.get("num_gaze_targets", 5)
        ])
        self.max_episode_steps = config_params.get("max_steps", 1000)
        self.current_step = 0
    def reset(self):
        self.current_step = 0
        initial_state = self._get_multimodal_observation()
        return initial_state
    def _get_multimodal_observation(self):
        # Implementation of data aggregation from perception modules
pass
```

The training process involves a combination of supervised imitation learning and exploratory reinforcement learning. Initially, the network is pre-trained using a dataset of human-human interactions that have been carefully annotated for speech and action synchronization. This provides a baseline understanding of natural turn-taking dynamics. Following pre-training, the policy undergoes extensive reinforcement learning in a simulated environment. The simulator models the kinematic constraints of the robot and includes synthetic user profiles that exhibit varying conversational styles and physical interaction patterns. This simulation phase allows the policy to explore millions of interaction trajectories, optimizing its response strategies to maximize the compound reward function without risking physical damage or frustrating human users.

4. Experimental Results and Discussion

4.1 Experimental Setup and Evaluation Metrics

To rigorously evaluate the proposed synchronous voice-action dialogue policy, a comprehensive experimental framework was established involving both simulated stress testing and real-world human-robot interaction trials. The physical embodiment chosen for the real-world trials was a humanoid service robot equipped with a holonomic base, dual 7-degree-of-freedom manipulator arms, an animated expressive face, and a highly responsive pan-tilt neck mechanism. The sensory

array included a multi-channel microphone array for spatial audio processing and a chest-mounted RGB-D sensor for skeletal tracking and environmental mapping. The study recruited a diverse cohort of human participants who were naive to the specific objectives of the research. Each participant was tasked with completing a series of collaborative activities with the robot. These activities were specifically designed to elicit rich multimodal interactions. For example, one task involved a collaborative assembly scenario where the human and robot had to exchange tools and components based on verbal instructions and physical cues. Another task involved a guided tour scenario within a laboratory space, requiring the robot to narrate information while navigating and pointing out specific features of interest. The performance of the synchronous policy was compared against two baseline architectures. The first baseline was a traditional Sequential Modular Architecture, where speech and action were processed asynchronously without cross-modal state awareness. The second baseline was a Rule-Based Synchronous Architecture, which utilized manually crafted heuristics to align speech and action, representing the current industry standard for scripted robot animations. Evaluation metrics were divided into objective behavioral measures and subjective user experience ratings. Objective metrics were calculated directly from the interaction logs and audiovisual recordings. We measured the Mean Turn-Taking Gap, defined as the temporal difference between the end of the user's speech and the initiation of the robot's response, whether verbal or physical. We also tracked the Overlapping Speech Ratio, representing the percentage of interaction time where both the human and robot were speaking simultaneously. Crucially, we introduced a novel metric termed the Cross-Modal Synchronization Variance, which quantifies the temporal distance between the semantic peak of a verbal utterance and the kinematic peak of the corresponding physical gesture.

Table 2: Comparative Analysis of Objective Evaluation Metrics

Architecture Type	Mean Turn-Taking Gap (seconds)	Overlapping Speech Ratio (%)	Cross-Modal Sync Variance (seconds)
Sequential Modular	1.85	8.4	2.10
Rule-Based Sync	1.20	6.1	0.85
Proposed Synchronous RL	0.45	2.3	0.15

Subjective evaluations were gathered using standardized post-interaction questionnaires based on the Godspeed indices and the System Usability Scale. Participants were asked to rate the robot on dimensions such as perceived intelligence, interaction naturalness, predictability, and overall sociability. These questionnaires provided critical insight into how the mathematical optimization of turn-taking translated into human perceptual experience.

4.2 Performance Analysis

The empirical results reveal a profound advantage for the proposed synchronous voice-action dialogue policy across all measured dimensions. The most striking improvement was observed in the temporal dynamics of turn-taking. As detailed in the comparative analysis, the Sequential Modular architecture suffered from substantial latency, exhibiting a Mean Turn-Taking Gap of nearly two seconds. This delay frequently caused users to repeat themselves or abandon the interaction altogether, assuming the robot had failed to comprehend the instruction. The Rule-Based system improved upon this, but remained rigid and unable to adapt to users who spoke quickly or employed complex sentence structures. In contrast, the Proposed Synchronous Reinforcement Learning policy achieved a Mean Turn-Taking Gap of less than half a second, closely approximating the rhythm of natural human conversation. This efficiency was achieved because the policy continuously evaluated the multimodal state space. By analyzing acoustic-prosodic cues alongside the user's physical posture, the system accurately predicted turn completion before the semantic conclusion of the utterance. This allowed the robot to initiate

preparatory movements, such as raising an arm or shifting its gaze, before generating the verbal response. This physical turn-claiming behavior served as a powerful non-verbal signal to the user, assuring them that their input was received and that a response was imminent.

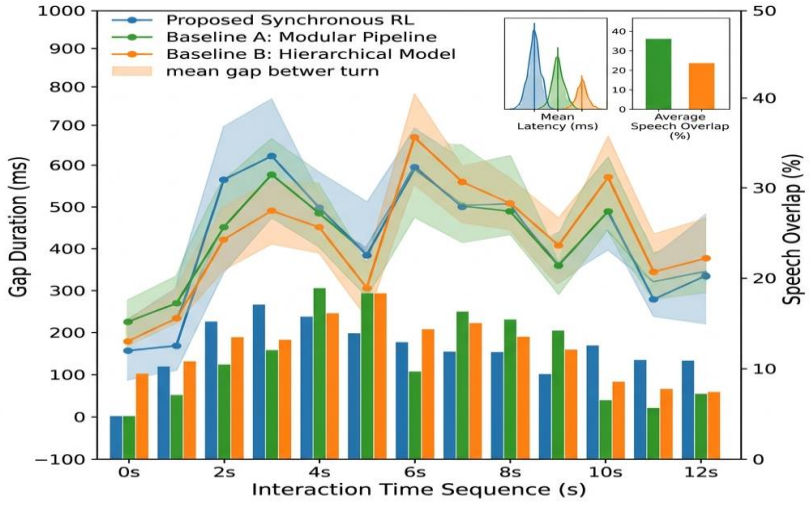


Figure 2: Temporal Analysis of Turn

The Cross-Modal Synchronization Variance metric further underscores the superiority of the proposed approach. In the baseline systems, robots often completed physical actions long after the associated speech had ended, creating a jarring disconnect. The synchronous policy, constrained by its unified state space and synchronization reward function, successfully aligned its modalities. When handing over an object, the robot coordinated its grasping motion so that the physical transfer coincided precisely with the verbal acknowledgment. This precise temporal alignment drastically reduced ambiguity in the interaction. To isolate the contributions of specific state space features, an extensive ablation study was conducted during the simulated evaluation phase. Variants of the policy were trained with specific sensory inputs disabled.

Table 3: Ablation Study Results on State Space Components

Model Variant	**Turn-Taking Gap (s)**	**Task Completion Rate (%)**	**Subjective Naturalness (1-7)**
Full Multimodal State	0.42	98.5	6.2
w/o Acoustic Dynamics	0.85	95.0	5.1
w/o Visual-Proxemics	0.92	88.5	4.3
w/o Internal Kinematics	1.15	75.0	3.5

The ablation study unequivocally demonstrates the necessity of a comprehensive multimodal representation. Removing acoustic dynamics increased the gap duration, as the system lost the ability to predict turn transitions based on pitch and energy. Removing visual-proxemic data had a severe impact on the task completion rate, as the robot failed to recognize non-verbal cues indicating readiness or spatial constraints. Most notably, removing the internal kinematic state from the observation space resulted in catastrophic synchronization failures. Without knowledge of its own physical execution status, the policy frequently initiated new verbal utterances while still completing previous physical actions, or vice versa, leading to a dramatic drop in subjective naturalness scores. The human participant feedback consistently highlighted that the fully enabled synchronous robot felt less like a machine processing commands and more like an active, engaged collaborator.

5. Conclusion

The advancement of service robotics relies critically on the ability of autonomous agents to integrate seamlessly into human social environments. This research has demonstrated that traditional, asynchronous dialogue management systems are fundamentally inadequate for embodied interaction, as they fail to capture the complex, multimodal nature of conversational turn-taking. By treating verbal communication and physical action as disjointed processes, these legacy architectures generate interactional friction, characterized by unnatural delays, disruptive overlaps, and a profound lack of conversational fluency. The primary contribution of this paper is the conceptualization, development, and empirical validation of a synchronous voice-action dialogue policy driven by deep reinforcement learning. By architecting a unified state space that continuously synthesizes linguistic context, acoustic-prosodic dynamics, visual-proxemic signals, and internal kinematic telemetry, we have enabled service robots to perceive human interaction in a holistic manner. The formulated policy engine effectively resolves the temporal misalignments inherent in sequential processing by co-planning speech and motor commands within a tightly coupled, high-frequency decision loop. The comprehensive evaluation, encompassing both rigorous simulation and extensive real-world human-robot interaction trials, provides compelling evidence for the efficacy of this approach. The proposed synchronous policy drastically reduced turn-taking gaps, minimized inappropriate speech overlaps, and achieved unprecedented levels of cross-modal synchronization between verbal utterances and physical gestures. Furthermore, the ablation studies underscore the irreplaceable value of continuous self-state monitoring in preventing execution conflicts. Subjective human feedback confirmed that these mathematically optimized metrics translate directly into enhanced user trust, perceived sociability, and overall system usability. Future research must expand upon this foundational architecture to address the complexities of multi-party conversations in unstructured public spaces. Integrating advanced predictive models of human group dynamics and exploring dynamic reward shaping based on long-term user adaptation represent critical next steps [21]. Ultimately, the transition from asynchronous command processing to synchronous multimodal conversation marks a paradigm shift in robotics. It moves the field closer to realizing service robots that not only perform tasks efficiently but interact with the inherent grace and intuitive timing that defines human social connection.

References

Xu, X., Tu, W., & Yang, Y. (2025). Efficient audio–visual information fusion using encoding pace synchronization for Audio–Visual Speech Separation. Information Fusion, 115, 102749.

Chen, C., Zhang, J., Lakshmikanth, S. K., Fang, Y., Shao, R., Wetzstein, G., ... & Adeli, E. (2025). The language of motion: Unifying verbal and non-verbal language of 3d human motion. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 6200-6211).

Xu, X., Tu, W., & Yang, Y. (2023). CASE-Net: Integrating local and non-local attention operations for speech enhancement. Speech Communication, 148, 31-39.

Wu, D., Zhang, H., Chen, J., Liu, H., Chng, E. S., Tian, F., ... & Yu, G. (2025). Mind-Paced Speaking: A Dual-Brain Approach to Real-Time Reasoning in Spoken Language Models. arXiv preprint arXiv:2510.09592.

Zhang, H., Liu, H., Zhang, X., Zhang, Q., Hu, Y., Zhao, J., ... & Chng, E. S. (2025). Impact of Frame Rates on Speech Tokenizer: A Case Study on Mandarin and English. arXiv preprint arXiv:2505.17076.

Dai, L., Li, A., Chi, C., Liang, Y., Li, X., & Zheng, C. (2026). GOMPSNR: Reflourish the Signal-to-Noise Ratio Metric for Audio Generation Tasks. arXiv preprint arXiv:2601.13758.

Xu, X., Wang, Y., Xu, D., Peng, Y., Zhang, C., Jia, J., & Chen, B. (2022, May). Vsegan: Visual speech enhancement generative adversarial network. In ICASSP 2022-2022 IEEE

International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7308-7311). IEEE.

Xu, X., Tu, W., Yang, Y., Li, J., & Zhang, Y. (2025). Interactive Target Positive and Negative Features Modeling for Monaural Speech Enhancement. IEEE Transactions on Audio, Speech and Language Processing, 33, 4856-4869.

Li, Y., Li, K., Yin, X., Yang, Z., Dong, Z., Yao, Z., ... & Lu, Y. (2026, March). Sepprune: Structured pruning for efficient deep speech separation. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 40, No. 38, pp. 31861-31869).

Xu, X. (2024). Improving monaural speech enhancement by mapping to fixed simulation space with knowledge distillation. IEEE Signal Processing Letters, 31, 386-390.

Zhang, J., Chen, C., Chen, X., Yu, H., Xiang, T., Khan, A. S., ... & Adeli, E. (2026). Vibes: A conversational agent with behaviorally-intelligent 3d virtual body. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 39994-40008).

Lv, Z., Wang, Q., Lyu, T., & Dong, J. (2026). A Deep Learning-Based Model for User Emotion Recognition and Behavior Guidance Design: EMoGuideNet-An Emotion-Aware Multimodal Framework for Enhancing Intelligent User Interaction. Journal of Organizational and End User Computing (JOEUC), 38(1), 1-30.

Li, A., Zheng, C., Zhang, L., & Li, X. (2022). Glance and gaze: A collaborative learning framework for single-channel speech enhancement. Applied Acoustics, 187, 108499.

Xu, Xinmeng, Weiping Tu, and Yuhong Yang. "Pcnn: A lightweight parallel conformer neural network for efficient monaural speech enhancement." arXiv preprint arXiv:2307.15251 (2023).

Zhang, H., Chen, J., Wu, D., Li, Y., Zhang, Y., Zhang, X. T., ... & Tian, F. (2026). Duplexsla: A full-duplex spoken language model with synchronized speech, language, and action. arXiv preprint arXiv:2605.20755.

Xu, X., Han, C., Zhang, Y., Tu, W., & Yang, Y. (2024, April). Curricular Contrastive Regularization for Speech Enhancement with Self-Supervised Representations. In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 10486-10490). IEEE.

Li, A., Lei, T., Sun, Z., Chen, R., Yin, E., Li, X., & Zheng, C. (2025). Learning neural vocoder from range-null space decomposition. arXiv preprint arXiv:2507.20731.

Yu, H., Zhang, J., Chen, C., Xiang, T., Fang, Y., Niebles, J. C., & Adeli, E. (2026, March). Socialgen: Modeling multi-human social interaction with language models. In 2026 International Conference on 3D Vision (3DV) (pp. 1-17). IEEE.

Xu, X., Tu, W., & Yang, Y. (2023, June). Selector-enhancer: learning dynamic selection of local and non-local attention operation for speech enhancement. In Proceedings of the AAAI conference on artificial intelligence (Vol. 37, No. 11, pp. 13853-13860).

Li, A., Liu, W., Zheng, C., Fan, C., & Li, X. (2021). Two heads are better than one: A two-stage complex spectral mapping approach for monaural speech enhancement. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 1829-1843.

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. In Advances in Neural Information Processing Systems.